IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10414-8

Colloquia: CSFI 2008

A HPC and Grid enabling framework for genetic linkage analysis of SNPs

L. Milanesi, A. Calabria, D. Di Pasquale, M. Gnocchi, A. Orro
and G. Trombetti

Istituto Tecnologie Biomediche, CNR - Via Fratelli Cervi 93, 20090 Segrate (MI), Italy

(ricevuto il 4 Agosto 2009; pubblicato online il 6 Ottobre 2009)

Summary. — Understanding the structure, function and development of the human genome is a key factor to improve the quality of life. In order to achieve this goal developing and using a modern ICT infrastructure is essential, and can exploit next generation High Performance Computing (HPC) systems beyond the Petaflop scale in a collaborative and efficient way. The genetic linkage analysis of Single Nucleotide Polymorphism (SNP) markers has recently become a very popular approach for genetic epidemiology and population studies, aiming to discover the genetic correlation in complex diseases. The high computational cost and memory requirements of the major algorithms proposed in the literature make analyses of medium/large data sets very hard on a single CPU. A Grid based facility has hence been set up upon a high-performance infrastructure, the EGEE Grid, in order to create a tool for achieving whole-genome linkage analysis.

PACS 87.85.mg – Genomics.
PACS 89.70.Eg – Computational complexity.

1. – Introduction

Genomics sequencing projects, neuroscience and new imaging technologies applied to modern medicine research are producing huge amounts of raw data. Biomedical and Neuroscience research laboratories are moving towards collaborative environments created through the sharing of resources, in which heterogeneous and dispersed health data, such as molecular data (*e.g.*, genomics, proteomics), cellular data (*e.g.*, pathways), tissue data, population data (*e.g.*, genotyping, SNP, epidemiology), as well the data generated by large scale analysis (*e.g.*, simulation data, modelling of organs and Brain) will need to be accessed by users worldwide in order to carry out large-scale analysis and research in multiple countries. In this scenario, a grid infrastructure appears a good fit as it can allow to store and manipulate extremely large amounts of heterogeneous data, and can allow the biomedical community members to have access to a more complete resource of information. In fact, the future trend for the biomedical scientific research is towards computing Grids, for data crunching applications, and data Grids for distributed storage of large amounts of accessible data and the provisioning of tools to all users. Biologists and medical researchers are becoming acquainted with the integration of wet-lab equipments with large infrastructures of computers for both daily lab-based experiments and large-scale experimental collaborations schemes.

Research studies in genetics of complex traits are used to study the independent and/or interactive contribution of genetic factors in human pathology, either with genome-wide approaches or with focused approaches on candidate genes or molecular pathways. In recent years, improvements in genotyping technologies analogous to those in the other biomedical sectors have greatly increased the amount of genetic data available, and made the study of DNA variations possible in large sample sizes, allowing more complex associations of genetic patterns to different clinical phenotypes. For the analysis of such huge amounts of data, however, complex bioinformatics workflows are used, and all data traditionally needed to be locally available during the computation. In this paper we present an approach for the distributed computation of large genetic SNP and mutation studies on high throughput genotyping data.

2. – Methods

Genetic Linkage Analysis is a statistical method for detecting genetic linkage between disease loci and markers of known locations by following their inheritance in families through the generations. This is a NP-hard problem [1] and the computational cost and memory requirements of the major algorithms proposed in the literature grows exponentially with pedigree size and markers' number. Implementations of the mentioned algorithms [2-4] reflect these limits making analyses of medium/large data sets very hard on a single CPU.

The aim of the present work is to enable the use of high-performance computing infrastructures, such as Clusters and Grid Infrastructures for the execution of linkage analysis pipeline on large data sets, especially for SNPs, dense biallelic markers (from 10K per chip to more than 1 million). In order to solve the computational problem and to exploit the intrinsic parallelism of the linkage analysis consisting in the independent comparison of markers against genotypes of each pedigree element, we created a specific heuristic which can compute different analyses with a subset of the markers. Although the heuristic gives lower joint results, the correctness of the proposed approach was tested setting up a benchmark analysis with a given dataset: we extracted 2230 SNPs from a chromosome and from these SNPs we formatted input files for the linkage analysis using different subsets of SNP for each single run: 30, 50 and 150 markers.

The distributed analysis pipeline is parameterized and monitored with the support of a web interface, designed to simplify and speed up the whole process of linkage analysis. Users can create the analysis pipeline arranging customizable modules representing pipeline steps; each module can be customized choosing different types of input files, introducing data pre-processing steps, selecting the algorithm and parameters, and choosing the HPC environment to be used. At the back-end level the pipeline engine splits the workload into small jobs and distributes analysis tasks over the available resources, either the Cluster nodes or the Grid computing elements; at the Cluster side the distribution of the jobs is natively managed by the pipeline engine, while at the Grid side this is achieved by VNAS [5]: a software framework built on top of the EGEE Grid middleware featuring enhanced submission modes for jobs, completion callbacks, automatic monitoring of jobs

Table I. – *Test results: data derived from different genotyping chips were analyzed with the Genehunter software using* 3 *computational infrastructures: our Grid-based system, a* 280 *cores Cluster and a single* 2 GHz *CPU Workstation. The left-hand side columns show the challenges characteristics, while on the right the challenges' durations are reported, expressed in hours, as resulted for the various execution environments.*

Genotype chip	Number	Computational cost (hours)		
(SNPs)	of jobs	1 CPU 2 GHz	Cluster	EGEE Grid
10 k	6	33	7	13
66 k	35	220	10	18
100 k	60	333	11	20
317 k	172	1056	13	36
370 k	206	1233	14	41
500 k	278	1665	16	48
670 k	373	2233	17	55
1 M	556	3332	20	61

over the Grid and automatic resubmission in case of failure, substantially guaranteeing successful completion of the computational jobs over the gLite Grid platform.

3. – Results

Tests were run to evaluate both the efficiency in terms of computational time and functionality of the proposed approach, obtaining an estimate of the EGEE Grid performances in comparison to a single 2 GHz CPU workstation and to the cluster composed by 280 CPU cores. The test, run using the Merlin software, involved the analysis of pedigrees composed by 32 subjects, including individuals genotyped using different genotyping chips having markers from 10k up to 1 million of SNPs each. The total number of runs needed to process all SNP data produced by each genotyping chip was split into jobs with an estimated duration of 6 hours each. Test results are summarized as follows: Table I shows the duration of the challenges in hours and fig. 1 displays the relative plot.

Comparing the results of different computation infrastructures we can see that distributed analysis pipelines with number of data (linkage variables) close to the computational limits for a mid-range workstation achieved significant improvements in computational time compared to dual-core 2 GHz CPU execution and considering markers chips greater than 100 k, the advantage of the distributed architecture gets proportionally bigger, due to the difference between linear increase of computational time for the sequential run on the single CPU and the saturation trend of the parallelized data flow obtained distributing the workload on the Grid's computing elements. For a standard linkage analysis based on 317k SNP chips, we achieved speedup of about 30 on Grid environment and up to 80 on cluster with respect to the single CPU sequential execution. The tests also show that the average performance of the EGEE Grid with the latest VNAS version can compete with a cluster environment in terms of pure computing time, where the difference is due to grid communication overhead.

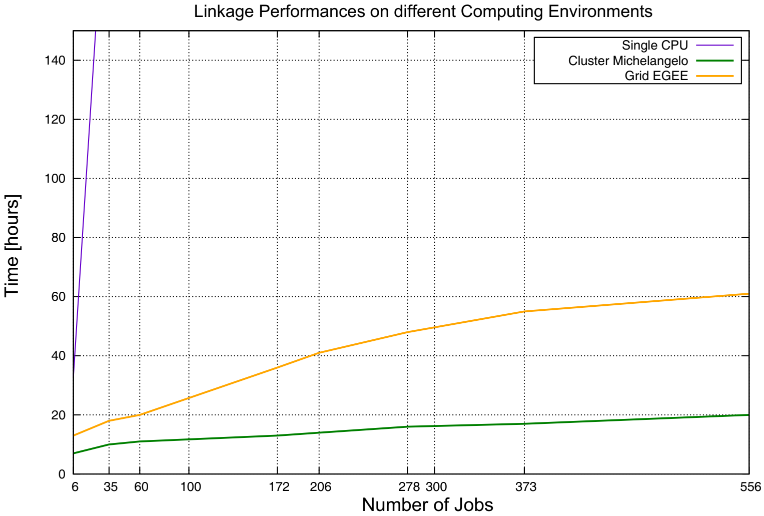

Fig. 1. – Performance test results: efficiency of the approach in the various environments.

4. – Conclusions

In this paper we described our effort at producing an application that allows scientists to easily launch genetic linkage analysis computations for medium/large challenges over a distributed computational infrastructure such as computational clusters and the EGEE Grid. The application provides easy accesses to high-performance and distributed computational platforms to users having only very basic knowledge in computer science. The computational pipeline offers customizable algorithms and parameters for genetic linkage analysis, features a friendly web user interface, achieves high performances using parallel processing tasks over a distributed environment, and is based upon VNAS, a framework managing low-level interactions and providing a layer of reliability and abstraction over distributed computing elements. The correctness of this approach has been tested by comparing the results obtained against earlier known and renowned non-distributed solutions.

* * *

This work has been supported by the NET2DRUG, EGEE-III, BBMRI, EDGE European projects, by the MIUR FIRB LITBIO (RBLA0332RH), ITALBIONET (RBPR05ZK2Z) Project.

REFERENCES

[1] Piccolboni A. and Gusfield D., *J. Comput. Biol.*, **10** (2003) 763.
[2] Elston R. C. and Stewart J., *Hum. Hered.*, **21** (1971) 523.
[3] Fishelson M. and Geiger D., *Bioinformatics*, **18** (2002) S189.
[4] Lander E. S. and Green P., *Proc. Nat. Acad. Sci. U. S. A.*, **84** (1987) 2363.
[5] Trombetti G. A., Bonnal R. J. P., Rizzi E., Bellis G. D. and Milanesi L., *BMC Bioinformatics*, **8** (2007) S22.